Econ 7801 (time series econometrics) study guide

Mckay Jensen

December 2019

1 Examples of time series models

One of the economic applications of stochastic difference equations that we talked about in class is based on a model presented in Samuelson's 1939 paper, "Interactions between the Multiplier Analysis and the Principle of Acceleration." The model that Samuelson proposed relates national income (y_t), consumption (c_t), private investment (i_t), and government spending (g_t) as:

$$\begin{aligned} y_t &= g_t + c_t + i_t \\ c_t &= \alpha y_{t-1} \\ i_t &= \beta(c_t - c_{t-1}) \end{aligned}$$

where $0 < \alpha < 1$ and $\beta > 0$.

The Enders book ignores government spending but adds a stochastic component to the consumption and investment difference equations:

$$\begin{aligned} y_t &= c_t + i_t \\ c_t &= \alpha y_{t-1} + \epsilon_{ct} \\ i_t &= \beta(c_t - c_{t-1}) + \epsilon_{it} \end{aligned}$$

ϵ_{ct} and ϵ_{it} are assumed to be white noise.

We can derive reduced-form equations from this system of structural difference equations. For example, putting y_t in terms of only its own lags and the stochastic components, we get

$$y_t = \alpha(1+\beta)y_{t-1} - \alpha\beta y_{t-2} + (1+\beta)\epsilon_{ct} + \epsilon_{it} - \beta\epsilon_{ct-1}$$

Given real-world data for y_t, we can run a regression using the above difference equation to estimate the parameters α and β and see how well this model explains real-world conditions.

Another application we discussed is using stochastic difference equations to test the "unbiased forward rate" hypothesis, i.e. the hypothesis that forward rates (f_t) should be

unbiased estimates of future spot exchange rates (s_{t+1}). If the hypothesis is true, we would expect the following relationship to hold:

$$s_{t+1} = f_t + \epsilon_{t+1}$$

(Again, ϵ_t is a white noise process.)

We can test this hypothesis by running a regression on the difference equation

$$s_{t+1} = \alpha_0 + \alpha_1 f_t + \epsilon_{t+1}.$$

If α_0 and α_1 do not differ from 0 and 1, respectively, on a statistically significant level, then we can maintain the unbiased forward rate hypothesis.

2 White noise and moving average processes

(a) A sequence of random variables $\{\epsilon_t\}$ is a white noise process if, for each t and for all lags $s = 0, 1, 2, \ldots$ and $j = 1, 2, 3, \ldots$, it satisfies the following conditions:

$$\begin{aligned} &\mathrm{E}[\epsilon_{t-s}] = 0 \\ &\mathrm{E}[\epsilon_{t-s}^2] = \sigma^2 \quad \text{for some constant } \sigma^2 > 0 \\ &\mathrm{E}[\epsilon_{t-s}\epsilon_{t-s-j}] = 0 \end{aligned}$$

That is, each term of the sequence must have zero mean, all terms must have the same variance, and distinct terms must have zero covariance with each other.

(b) A moving average process of order q is a sequence $\{x_t\}$ where, for some constants $\beta_0, \beta_1, \ldots, \beta_q$ and some white noise process $\{\epsilon_t\}$,

$$x_t = \sum_{i=0}^{q} \beta_i \epsilon_{t-i} \quad \text{for all } t.$$

(c) Given a white noise process $\{\epsilon_t\}$, let $\{x_t\}$ be the moving average process defined by

$$x_t = \epsilon_t + \beta_1 \epsilon_{t-1},$$

where $\beta_1 \neq 0$.

For all $s = 0, 1, 2, \ldots$ and $j = 1, 2, 3, \ldots$, we have (based off the properties of expected value of $\{\epsilon_t\}$)

$$\begin{aligned} \mathrm{E}[x_{t-s}] &= \mathrm{E}[\epsilon_{t-s} + \beta_1 \epsilon_{t-s-1}] \\ &= \mathrm{E}[\epsilon_{t-s}] + \beta_1 \mathrm{E}[\epsilon_{t-s-1}] \\ &= 0 + \beta_1 \cdot 0 = 0 \end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}[x_{t-s}^2] &= \mathrm{E}[(\epsilon_{t-s} + \beta_1 \epsilon_{t-s-1})^2] \\
&= \mathrm{E}[\epsilon_{t-s}^2 + 2\beta_1 \epsilon_{t-s}\epsilon_{t-s-1} + \beta_1^2 \epsilon_{t-s-1}^2] \\
&= \mathrm{E}[\epsilon_{t-s}^2] + 2\beta_1 \mathrm{E}[\epsilon_{t-s}\epsilon_{t-s-1}] + \beta_1^2 \mathrm{E}[\epsilon_{t-s-1}^2] \\
&= \sigma^2 + 2\beta_1 \cdot 0 + \beta_1^2 \sigma^2 = (1 + \beta_1^2)\sigma^2
\end{aligned}$$

but

$$\begin{aligned}
\mathrm{E}[x_{t-s}x_{t-s-j}] &= \mathrm{E}[(\epsilon_{t-s} + \beta_1 \epsilon_{t-s-1})(\epsilon_{t-s-j} + \beta_1 \epsilon_{t-s-j-1})] \\
&= \mathrm{E}[\epsilon_{t-s}\epsilon_{t-s-j} + \beta_1 \epsilon_{t-s}\epsilon_{t-s-j-1} \\
&\qquad + \beta_1 \epsilon_{t-s-1}\epsilon_{t-s-j} + \beta_1^2 \epsilon_{t-s-1}\epsilon_{t-s-j-1}] \\
&= \mathrm{E}[\epsilon_{t-s}\epsilon_{t-s-j}] + \beta_1 \mathrm{E}[\epsilon_{t-s}\epsilon_{t-s-j-1}] \\
&\qquad + \beta_1 \mathrm{E}[\epsilon_{t-s-1}\epsilon_{t-s-j}] + \beta_1^2 \mathrm{E}[\epsilon_{t-s-1}\epsilon_{t-s-j-1}]
\end{aligned}$$

so when $j = 1$,

$$\begin{aligned}
\mathrm{E}[x_{t-s}x_{t-s-1}] &= \mathrm{E}[\epsilon_{t-s}\epsilon_{t-s-1}] + \beta_1 \mathrm{E}[\epsilon_{t-s}\epsilon_{t-s-2}] + \beta_1 \mathrm{E}[\epsilon_{t-s-1}^2] \\
&\qquad + \beta_1^2 \mathrm{E}[\epsilon_{t-s-1}\epsilon_{t-s-2}] \\
&= 0 + \beta_1 \cdot 0 + \beta_1 \sigma^2 + \beta_1^2 \cdot 0 \\
&= \beta_1 \sigma^2 \neq 0
\end{aligned}$$

(Here, $\sigma^2 > 0$ is the variance of the $\{\epsilon_t\}$ process.)

Subsequent terms have nonzero covariance, so $\{x_t\}$ is not a white noise process.

(d) Let $\{y_t\}$ be a sequence of random variables. $\{y_t\}$ is called a covariance stationary process if, for each t and for all lags $s = 0, 1, 2, \ldots$ and $j = 0, 1, 2, \ldots$, there are constants μ, $\sigma_y^2 > 0$, and $\gamma_s \geq 0$ (with a different γ_s for each s), such that

$$\begin{aligned}
&\mathrm{E}[y_{t-s}] = \mu \\
&\mathrm{E}[(y_{t-s} - \mu)^2] = \sigma_y^2 \\
&\mathrm{E}[(y_{t-j} - \mu)(y_{t-j-s} - \mu)] = \gamma_s
\end{aligned}$$

That is, all terms of the series have the same mean and variance, and pairs of terms separated by the same lag have the same covariance. (Notice that $\gamma_0 = \sigma_y^2$, so technically the second condition is redundant.)

(e) A time series being *stationary* essentially means that it has no trend over time, or, a bit more precisely, its statistical properties do not change over time. We focused

above on *covariance stationarity*, also called "weak" or "wide-sense" stationarity. There is a stronger type of stationarity, "strict" or "strong" stationarity:

A time series $\{y_t\}$ is *strictly stationary* if, for any sequence of integers $t_1, t_2, \ldots, t_k$ and shift s, the joint probability distributions of

$$(y_{t_1}, y_{t_2}, \ldots, y_{t_k}) \text{ and } (y_{t_1+s}, y_{t_2+s}, \ldots, y_{t_k+s})$$

are the same.

Note that if a time series is strictly stationary, it is also covariance stationary.

Ergodicity is a rather more complicated property: roughly speaking, a series is ergodic if its statistical properties can be deduced from a single, sufficiently long, random sample of the process. The formal definition for ergodicity is complicated, but there is a useful equivalent definition that is a bit more intuitive:

A sequence $\{y_t\}$ is an *ergodic stochastic process* if and only if for any shift $\tau_1, \ldots, \tau_k$ and function $g : R^{k+1} \to R$, we have

$$\frac{1}{n}\sum_{t=1}^{n} g(y_t, y_{t+\tau_1}, \ldots, y_{t+\tau_k}) \overset{a.s.}{\to} \mathrm{E}[g(y_0, \ldots, y_{t+\tau_k})].$$

(That definition is based on Subba Rao, *A Course In Time Series Analysis*, 2018.)

You can see how testing for ergodicity involves looking at the behavior of the series over time (i.e. looking at a single instantiation of the series and improving estimates of its parameters by looking at longer stretches of time), while testing for stationarity involves estimating the parameters (for covariance stationarity, the mean, variance, and covariances) of the constituent random variables by taking samples from their theoretical population.

We should note that ergodicity implies stationarity, but stationarity does not imply ergodicity.

Interestingly enough, a sufficient condition for a covariance-stationary process with autocovariances $\{\gamma_s : s = 0, 1, 2, \ldots\}$ to be ergodic is for the sum $\sum_{s=0}^{\infty} |\gamma_s|$ to be finite.

3 Stability of ARMA processes

(a) A series is *stable* if it does not diverge to $\pm\infty$. Let's consider an ARMA model

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=0}^{q} \beta_i \epsilon_{t-i}.$$

The MA part can't grow uncontrollably since $\{\epsilon_t\}$ is assumed to have mean zero, so we focus on the AR part: from basic theory of linear difference equations, the difference equation

$$y_t = \sum_{i=1}^{p} a_i y_{t-i}$$

is stable if and only if the roots of the characteristic polynomial

$$\rho(\xi) = \xi^p - \sum_{i=1}^{p} a_i \xi^{p-i}$$

are strictly within the unit circle in the complex plane. (An equivalent condition is that the "reverse characteristic polynomial" has all roots greater than one, but we'll stick with this condition for now because it works pretty much the same.)

Stationarity is the condition that we've been discussing above, basically that there is no trend over time. Note that this is quite a different sort of condition from stability; however, the two conditions are related. Specifically, necessary requirements for an ARMA(p, q) process $\{y_t\}$ to be covariance stationary are that q must be finite and that the MA part must be stable, i.e. all the roots of the characteristic polynomial ρ must have magnitude less than one.

(b) Let $\{y_t\}$ be the stochastic process defined by

$$y_t = a_0 + a_1 y_{t-1} + \epsilon_t$$

The characteristic polynomial for this process is

$$\rho(\xi) = \xi - a_1.$$

The only root of ρ is $\xi = a_1$. Therefore, the process is stable only if $|a_1| < 1$. Based on the previous discussion, $|a_1| < 1$ is also a condition for the process to be stationary.

4 Maximum likelihood estimation

(a) Suppose we have some statistical model (in the context of this class, an ARMA model) with parameters $\boldsymbol{\theta}$. For an ARMA(p, q) model,

$$\boldsymbol{\theta} = (a_0, a_1, \ldots, a_p, \beta_1, \beta_2, \ldots, \beta_q, \sigma_2).$$

If we have observed a sample $\boldsymbol{x}$ of size T,

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$$

then we can consider the joint probability density function of these data, $f(\boldsymbol{x};\boldsymbol{\theta})$, which we can think of as expressing the likelihood of seeing $\boldsymbol{x}$ given parameters $\boldsymbol{\theta}$. The likelihood function $L(\boldsymbol{\theta}|\boldsymbol{x})$ is that joint density treated as a function of $\boldsymbol{\theta}$, given the observed $\boldsymbol{x}$. The maximum likelihood estimator is

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\boldsymbol{\theta}|\boldsymbol{x})$$

(Note that it's sometimes easier to work with the log-likelihood $l(\boldsymbol{\theta}|\boldsymbol{x}) = \ln(L(\boldsymbol{\theta}|\boldsymbol{x})$ as they'll be maximized in the same place since ln is a monotonically increasing function.)

In order for this to work, we of course need to know what type of distribution the data is supposed to have. If we assume that $\boldsymbol{x}$ is distributed on a multivariate normal distribution, then the likelihood function is

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{x}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{x}\right)}{\sqrt{(2\pi)^T|\Sigma(\boldsymbol{\theta})|}},$$

where $\Sigma(\boldsymbol{\theta})$ is the covariance matrix as a function of $\boldsymbol{\theta}$. The log-likelihood function is then

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = -\frac{1}{2}\left(T\ln(2\pi) + \ln|\Sigma(\boldsymbol{\theta})| + \boldsymbol{x}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{x}\right),$$

which is considerably easier to work with.

5 ACF and PACF

The autocorrelation function (ACF) is just the autocorrelations as a function of the lag size, i.e., for a series $\{y_t\}$,

$$\text{ACF}(s) = \rho(y_t, y_{t-s}) = \gamma_s/\gamma_0.$$

The partial autocorrelation function (PACF) is similar, but eliminates the effects of intervening values. That is, if the correlation between y_t and y_{t-s} is partially or wholly due to both y_t and y_{t-s} being correlated with some y_{t-j} $(0 < j < s)$, then that effect will not be part of the partial autocorrelation between y_t and y_{t-s}.

The ACF and the PACF can be used to identify what sort of process you're dealing with. The ACF of an ARMA(p, q) process will begin to decay after lag q, and the PACF of an ARMA(p, q) process will begin to decay after lag p. The rate and possible oscillatory behavior of the decay is determined by the AR and MA parts, respectively.

6 Ljung-Box Q-statistic

The Ljung-Box Q-statistic can be used to determine if a group $\{r_k : k = 1, 2, \ldots, s\}$ of sample autocorrelations is significantly different from 0. The statistic is

$$Q = T(T+2)\sum_{k=1}^{s}\frac{r_k^2}{T-k},$$

with T being the number of terms sampled from the underlying time series. Under the null hypothesis that $r_k = 0$ for all k, Q is $\chi^2(s)$ distributed.

The Q-statistic can also be used to check if the residuals of an estimated time series model behave like white noise. The only difference is that when the model is estimated, the degrees of freedom are reduced by the number of estimated coefficients. So if we want to check that the residuals of an estimated ARMA(p, q) model behave like white noise, we would compute the Q-statistic for the autocorrelations of the residuals and check if it exceeds the critical value for a $\chi^2(s - p - q)$ distribution. (If the model has a constant term, it would be a $\chi^2(s - p - q - 1)$ distribution.)

7 Invertibility of ARMA processes

A process $\{y_t\}$ is invertible if it can be represented by a finite-order or convergent autoregressive process. In order for an ARMA model to have this property, the polynomial

$$1 + \sum_{i=1}^{q}\beta_i L^i$$

must have all roots outside the unit circle.

The process defined by $y_t = \epsilon_t + \beta_1\epsilon_{t-1}$ will be invertible if the roots of $1 + \beta_1 L$ are outside the unit circle. The only root is $L = -1/\beta_1$, so we need $|-1/\beta_1| > 1$, or, equivalently, $|\beta_1| < 1$.

8 Information criteria

The Akaike Information Criterion (AIC) is defined as

$$AIC = T\ln(\text{sum of squared residuals}) + 2n,$$

where T is the number of data points used to estimate the model, and n is the number of parameters estimated ($n = p + q$, or $n = p + q + 1$ if a constant term is included).

The Schwarz Bayesian Information Criterion (SBC) is defined as

$$SBC = T\ln(\text{sum of squared residuals}) + n\ln T.$$

Both of these criteria will move toward $-\infty$ as the fit of the model improves. However, since $n \ln T > 2n$ for $T > e^2$, the SBC penalizes the addition of unnecessary regression coefficients more than the AIC does. It's generally desirable not to add unnecessary coefficients (this helps us get a more "parsimonious" model), so the SBC is good in that respect, but we do need to make sure to check that the residuals appear to be white noise (which, if not true, would indicate that we're probably missing something). If using the AIC to evaluate a model, we should make sure that the t-statistic of each coefficient is at a significant level.

9 Forecasts from ARMA models

(a) We consider the model defined by $y_t = a_0 + a_1 y_{t-1} + \epsilon_t$

(a) The $s-$step ahead forecast is the expected value of y_{t+s} given the information available at time t. That is,

$$\mathrm{E}_t y_{t+s} = \mathrm{E}[y_{t+s} | y_t, y_{t-1}, y_{t-2}, \dots]$$

The one-step ahead forecast is

$$\mathrm{E}_t y_{t+1} = \mathrm{E}_t[a_0 + a_1 y_t + \epsilon_t] = a_0 + a_1 y_t$$

The two-step ahead forecast is

$$\begin{aligned} \mathrm{E}_t y_{t+2} &= \mathrm{E}_t[a_0 + a_1 y_{t+1} + \epsilon_t] \\ &= a_0 + a_1 \mathrm{E}_t y_{t+1} \\ &= a_0 + a_1(a_0 + a_1 y_t) \\ &= a_0 + a_0 a_1 + a_1^2 y_t \end{aligned}$$

The three-step ahead forecast is

$$\begin{aligned} \mathrm{E}_t y_{t+3} &= \mathrm{E}_t[a_0 + a_1 y_{t+2} + \epsilon_t] \\ &= a_0 + a_1 \mathrm{E}_t y_{t+2} \\ &= a_0 + a_1(a_0 + a_0 a_1 + a_1^2 y_t) \\ &= a_0 + a_0 a_1 + a_0 a_1^2 + a_1^3 y_t \end{aligned}$$

(b) The s-step ahead forecast error is the difference between the actual value of the process s steps ahead and the s-step ahead forecast. Formally,

$$y_{t+s} - \mathrm{E}_t y_{t+s}$$

The one-step ahead forecast error is

$$y_{t+1} - \mathrm{E}_t y_{t+1} = (a_0 + a_1 y_t + \epsilon_{t+1}) - E_t y_{t+1} = \epsilon_{t+1}$$

The two-step ahead forecast error is

$$\begin{aligned} y_{t+2} - \mathrm{E}_t y_{t+2} &= (a_0 + a_1 y_{t+1} + \epsilon_{t+2}) - E_t y_{t+2} \\ &= [a_0 + a_1(a_0 + a_1 y_t + \epsilon_{t+1}) + \epsilon_{t+2}] - E_t y_{t+2} \\ &= a_0 + a_0 a_1 + a_1^2 y_t + a_1 \epsilon_{t+1} + \epsilon_{t+2} - E_t y_{t+2} \\ &= a_1 \epsilon_{t+1} + \epsilon_{t+2} \end{aligned}$$

The three-step ahead forecast error is

$$\begin{aligned} y_{t+3} - \mathrm{E}_t y_{t+3} &= (a_0 + a_1 y_{t+2} + \epsilon_{t+3}) - E_t y_{t+3} \\ &= [a_0 + a_1(a_0 + a_1 y_{t+1} + \epsilon_{t+2}) + \epsilon_{t+3}] - E_t y_{t+3} \\ &= a_0 + a_0 a_1 + a_1^2(a_0 + a_1 y_t + \epsilon_{t+1}) + a_1 \epsilon_{t+2} + \epsilon_{t+3} - E_t y_{t+3} \\ &= a_0 + a_0 a_1 + a_0 a_1^2 + a_1^3 y_t + a_1^2 \epsilon_{t+1} + a_1 \epsilon_{t+2} + \epsilon_{t+3} - E_t y_{t+3} \\ &= a_1^2 \epsilon_{t+1} + a_1 \epsilon_{t+2} + \epsilon_{t+3} \end{aligned}$$

(b) You can get an idea of how good the forecasts of a model are just by looking at the variances of the forecast errors. For example, the variance of the three-step ahead forecast error above is

$$\mathrm{Var}(y_{t+3} - \mathrm{E}_t y_{t+3}) = (a_1^4 + a_1^2 + 1)\sigma^2.$$

Ideally, we'd like low variance since that makes it less likely that we'll get a very high forecast error.

Another way to test the forecasts a model makes is to estimate its parameters using only some of the available time data, use that estimated model to create forecasts, and then compare those forecasts with the actual future values. We can quantify the forecast accuracy with the "mean square prediction error" (MSPE). Suppose that we construct H one-step ahead forecasts and compute their errors (i.e. by estimating models on a partial set of the data, predicting the next value, computing the error, re-estimating with the next data point included, and so on). Denoting the ith error as e_i, the MSPE is

$$\mathrm{MSPE} = \frac{1}{H} \sum_{i=1}^{H} e_i^2$$

A lower MSPE suggests that a model makes better forecasts. This of course is really only helpful in comparison to some other model estimated on the same time series, which brings us to the next question...

(c) Suppose we compute H forecast errors for two models, which we'll call model 1 and model 2. If the errors are e_{1i} and e_{2i} $(i = 1, 2, \ldots, H)$, we can compute MSPEs for both:

$$\text{MSPE}_1 = \frac{1}{H}\sum_{i=1}^{H} e_{1i}^2$$

$$\text{MSPE}_2 = \frac{1}{H}\sum_{i=1}^{H} e_{2i}^2$$

If MSPE_1 is greater than MSPE_2 we would prefer model 1 to model 2 in terms of forecasting ability. However, in order to determine that the difference is statistically significant, we use the F-statistic

$$F = \frac{\text{MSPE}_1}{\text{MSPE}_2}$$

which, under then null hypothesis that $\text{MSPE}_1 = \text{MSPE}_2$, follows an $F(H, H)$ distribution (technically, we also have to satisfy the assumptions that the forecast errors have zero mean, are normally distributed, and are serially and contemporaneously uncorrelated). We can then conclude that the MSPEs differ on a statistically significant level if the F-statistic exceeds the relevant critical value.

10 Testing for structural breaks

Suppose that we suspect a structural break at a time t_m. One way to test this is to estimate two models based on the data before time t_m and the data after t_m. To test the hypothesis that the coefficients for the model in the first period are the same as the coefficients for the second, we compute the residual sum of squares for both, which we'll denote RSS_1 and RSS_2. We can then use an F-test with

$$F = \frac{[RSS - (RSS_1 + RSS_2)]/n}{(RSS_1 + RSS_2)/(T - 2n)}.$$

The degrees of freedom are n and $T - 2n$, with T being the total number of data points and n being the number of coefficients estimated. A higher F-statistic would indicate that it is less likely that $RSS = RSS_1 + RSS_2$ (which is what we would expect if all coefficients were equal), so a high F-statistic is evidence of a structural break at t_m.

We can't always assume that a break occurs at a neatly specified point in time. To test for more general "parameter instability," we can successively estimate the model using more and more of the time series, making one-step ahead forecasts at each step and computing their forecast errors. We can then compute a cumulative sum of errors statistic

$$CUSUM_N = \sum_{i=n}^{N} \frac{e_i(1)}{\sigma_e}, \quad N = n, \ldots, T-1$$

Here n is the first time point we estimate the model for, $e_i(1)$ is the one-step ahead forecast error at time i, and σ_e is the estimated standard deviation of the forecast errors. At the 5% significance level, each value of $CUSUM_N$ should be within a band of

$$\pm 0.948[(T-n)0.5 + 2(N-n)(T-n) - 0.5].$$

Otherwise, the time series likely displays parameter instability.

11 Trend-stationary and difference-stationary processes

(a) Neither trend-stationary nor difference-stationary processes are stationary – that is, they both have some sort of trend, and we need to remove that trend in order to examine the underlying stationary process. The difference is that trend-stationary processes have a *deterministic trend*, and difference-stationary processes have a *stochastic trend*. This means that the effects of shocks in trend-stationary processes are eventually eliminated (the time series will revert to the trend in the long run), while the effects of shocks persist in difference-stationary series. Formally, we can express a trend-stationary process $\{x_t\}$ as

$$x_t = \mu_t + x'_t,$$

where μ_t is a deterministic trend, and x'_t is a stationary process with zero mean. Therefore, in order to detrend a trend-stationary process, we just need to remove the deterministic μ_t part. For example, if we suspect that μ_t takes the form

$$\mu_t = \alpha_0 + \alpha_1 t$$

for some constants α_0 and α_1, then we could run the regression

$$x_t = \alpha_0 + \alpha_1 t + x'_t$$

with the residuals $\{x'_t\}$ being our detrended (stationary) series.

On the other hand, in the context of linear stochastic processes, a process

$$y_t = a_0 + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=0}^{q} b_i \epsilon_{t-i}$$

is difference-stationary if it has a "unit root," i.e. if at least one of the roots of its characteristic polynomial

$$\rho(\xi) = \xi^p - \sum_{i=1}^{p} a_i \xi^{p-i}$$

has magnitude equal to 1. For example, if $\{y_t\}$ is a random walk

$$y_t = y_{t-1} + \epsilon_t,$$

then we have $\rho(\xi) = \xi - 1$, which has root $\xi = 1$. Thus $\{y_t\}$ is a unit root (difference-stationary) process. To remove the trend from a difference-stationary process, we can take the dth difference of the process, with d depending on how many unit roots the process has. In general, if the dth difference of $\{y_t\}$, $\Delta^d y_t = (1 - L)^d y_t$, is a stationary ARMA$(p, q)$ process, then we call $\{y_t\}$ an ARIMA(p, d, q) process. In the case of y_t being a random walk (as above), we have

$$\Delta y_t = y_t - y_{t-1} = \epsilon_t,$$

so a random walk is an ARIMA$(0, 1, 0)$ process.

(b) It is important to use the correct technique to obtain a stationary process from a non-stationary process. As explained above, to remove the trend from a trend-stationary process, we estimate the deterministic trend (with a regression) and then simply subtract it out; to remove the trend from a difference-stationary time series with d unit roots, we take the dth difference of the series.

If we attempt to remove the trend from a trend-stationary process by taking a difference, we'll end up with a non-invertible unit root process, which just makes things even more complicated than before. For example, if we have the [trend-stationary] process

$$x_t = x_0 + a_1 t + \epsilon_t,$$

then the first difference is

$$\Delta x_t = a_1 + \epsilon_t - \epsilon_{t-1}.$$

The MA part of Δx_t has a unit root, so Δx_t is not invertible; attempting to detrend $\{x_t\}$ by taking differences is not the right way to go.

Likewise, attempting to remove the trend from a difference-stationary process by removing a deterministic trend component won't work because the trend in a difference-stationary series is stochastic, not deterministic. Modeling the stochastic trend as deterministic will yield misleading results that depend on the particular instantiation of the series being observed.

(c) A **random walk** has the form

$$y_t = y_{t-1} + \epsilon_t,$$

as in the example given in part (a). The general solution is

$$y_t = y_0 + \sum_{i=1}^{t} \epsilon_i.$$

As shown in part (a), a random walk has a unit root and is therefore difference-stationary. You can also see that it is not covariance-stationary rather easily by looking at the general solution: although $\mathrm{E}[y_t] = y_0$ for all t, the variance depends on the given value of t, with $\mathrm{Var}(y_t) = \sum_{i=1}^{t} \mathrm{Var}(\epsilon_i)$.

A **random walk with drift** has the form

$$y_t = a_0 + y_{t-1} + \epsilon_t, \quad a_0 \neq 0.$$

The general solution is

$$y_t = y_0 + a_0 t + \sum_{i=1}^{t} \epsilon_i.$$

Notice that here we have *both* a linear deterministic trend ($a_0 t$, "drift") and a stochastic trend ($\sum_{i=1}^{t} \epsilon_i$). We can get a stationary series with mean a_0 by taking the first difference:

$$\Delta y_t = a_0 + \epsilon_t.$$

As described in part (a), a **unit root process** is a process where at least one of the roots of the characteristic polynomial has magnitude 1. Both a random walk and a random walk with drift are examples of a unit root process (as they have the same characteristic polynomial). That is, any time series of the form

$$y_t = a_0 + y_{t-1} + \epsilon_t, \quad a_0 \in R$$

will be a unit root process.

Unit root processes are difference-stationary, so the proper way to remove their trend is to take differences of the series, with the number of differences to take equal to the number of unit roots. Consider, for example, the process defined by

$$y_t = a_0 + 2y_{t-1} - y_{t-2} + \epsilon_t$$

The characteristic polynomial is

$$\rho(\xi) = \xi^2 - 2\xi + 1$$

which has root 1 with multiplicity 2. So the process is a unit root process, integrated of order 2. Sure enough, the second difference is

$$\Delta^2 y_t = (1 - L)^2 y_t = y_t - 2y_{t-1} + y_{t-2}$$

so we have

$$y_t = 2y_{t-1} - y_{t-2} + \Delta^2 y_t$$

meaning it must be that

$$\Delta^2 y_t = a_0 + \epsilon_t,$$

which is, of course, stationary.

12 Unit root tests

(a) Suppose that we have time series data $\{y_t : t = 1, 2, \ldots, T\}$ and want to determine whether $\{y_t\}$ is a unit root process.

One way to test the hypothesis that $\{y_t\}$ comes from a process with a unit root is with the **augmented Dickey-Fuller (ADF) test**. The test consists of estimating the coefficients $\gamma, \delta_1, \delta_2, \ldots, \delta_{p-1}$ in the regression equation

$$\Delta y_t = \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + e_t.$$

Intuitively, if $\{y_t\}$ is an ARIMA$(p, 1, q)$ process, i.e. it has one unit root, then the lags $\Delta y_{t-1}, \Delta y_{t-2}, \ldots, \Delta y_{t-p+1}$ should provide an unbiased estimate of Δy_t, so knowing y_{t-1} should provide no additional useful information, and γ should be 0. Conversely, if $\{y_t\}$ has no unit root, then it is stationary, so it should exhibit regression to the mean and thus y_{t-1} should be negatively correlated with Δy_t. Thus if the estimated γ coefficient is significantly different from 0, then we can reject the null hypothesis of a unit root and conclude that $\{y_t\}$ is not a unit root process.

We can slightly modify the test to allow for the possibility of drift and of a deterministic linear trend, using the following regression equation:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + e_t.$$

If α and β are fixed at 0, then the test is the same as before; if we fix only β at 0, we allow for the possibility of drift; otherwise, we allow for both drift and a deterministic linear trend. In any case, we test the null hypothesis that $\gamma = 0$.

It should be noted that γ under $H_0 : \gamma = 0$ does *not* come from a t-distribution; the special "Dickey-Fuller distribution" that γ comes from can be generated via Monte Carlo methods. At any rate, the hypothesis test is, at a basic level, standard: if γ exceeds the critical value for its null distribution, we reject the null hypothesis and conclude that the process does not have a unit root.

Another caveat: the ADF test requires knowing beforehand the correct autoregressive order p; one way to find the right p is simply to start with a large value of p and continue reducing p until all the δ coefficients in the regression equation are statistically significant. Another way is to choose the p that minimizes the Schwarz Bayesian criterion (or some other similar criterion).

The ADF test is not the only test available to test for a unit root. Another test used relatively often is the **Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test**. The basic idea behind the KPSS test is to decompose $\{y_t\}$ into a deterministic trend, a random walk component, and white noise:

$$y_t = a_0 + \beta t + r_t + \epsilon_t$$

with

$$r_t = r_{t-1} + \mu_t,$$

where both $\{\epsilon_t\}$ and $\{\mu_t\}$ are white noise, with the variance of the latter being σ^2. Here, the null hypothesis is that $\{y_t\}$ is stationary (does *not* have a unit root), i.e., that $\beta = \sigma^2 = 0$. (We can also test for trend-stationarity by testing the null hypothesis that just σ^2 is 0.)

(b) The power of unit root tests can be quite low. That is, they can often fail to reject a false null hypothesis – in the case of the ADF test, the test frequently fails to reject the hypothesis of a unit root, especially when γ is close to zero. This is in part due to the fact that, in a finite sample, a trend-stationary process can be arbitrarily well approximated by a unit root (difference-stationary) process, and vice versa. One way to ensure that the power of the ADF test is as high as possible is to make sure that the model is correctly specified so that no more regressors are included than necessary; that is, we only want to include regressors that are part of the actual data-generating process. This means choosing the proper value for p (as discussed above in part (d)) and only including the coefficients α and β if there is reason to believe that the process includes drift and/or a linear trend.

Another technique that may increase the power of the ADF test is to first detrend the time series using generalized least squares, then run the ADF test (without α and β) on the detrended time series – this is referred to as the augmented Dickey-Fuller generalized least squares (ADF-GLS) test.

13 Estimation of nonstationary models

Suppose we have data for two time series, $\{y_t\}$ and $\{x_t\}$, and we want to estimate the coefficients for the regression equation

$$y_t = a_0 + a_1 x_t + e_t.$$

There are a few different cases we need to consider:

1. *Both $\{y_t\}$ and $\{x_t\}$ are stationary.* In this case, the classic regression model is appropriate, i.e. we can use OLS to estimate a_0 and a_1, and the residuals $\{e_t\}$ should be a stationary series.

2. *The $\{y_t\}$ and $\{x_t\}$ sequences are integrated of different orders.* For example, suppose that the two sequences are defined by

$$y_t = y_{t-1} + \epsilon_{yt}$$

$$x_t = \alpha x_{t-1} + \epsilon_{xt}, \quad |\alpha| < 1.$$

In this case $\{y_t\}$ is integrated of order 1 while $\{x_t\}$ is stationary. Assuming that $x_0 = y_0 = 0$, the residuals $\{e_t\}$ will take the form

$$e_t = \sum_{i=0}^{\infty} \epsilon_{y(t-i)} - a_1 \sum_{i=0}^{\infty} \alpha^i \epsilon_{x(t-i)}.$$

Although the series on the right is convergent, the one on the left is not and thus represents a stochastic trend component; so the sequence of residuals is not stationary. This violates classical assumptions for linear regression analysis, so results from the proposed regression equation are meaningless (we may see correlative relationships where none exist).

3. *Both $\{y_t\}$ and $\{x_t\}$ are non-stationary and integrated of the same order, and sequence of residuals $\{e_t\}$ contains a stochastic trend (is non-stationary).* This is the classic case of the spurious regression. For example, let's assume that $\{y_t\}$ and $\{x_t\}$ are both random walks:

$$y_t = y_{t-1} + \epsilon_{yt}$$

$$x_t = x_{t-1} + \epsilon_{xt}$$

Both time series are integrated of order one, and the residuals $\{e_t\}$ will take the form

$$e_t = \sum_{i=0}^{\infty} \epsilon_{y(t-i)} - a_1 \sum_{i=0}^{\infty} \epsilon_{x(t-i)},$$

which clearly exhibits a stochastic trend and is thus not stationary. As in the previous case, regression results will be meaningless. However, regressing on first differences (or, in general, dth differences for processes integrated of order d), may be appropriate; for the case where $\{y_t\}$ and $\{x_t\}$ are $I(1)$, each of $\{\Delta y_t\}$, $\{\Delta x_t\}$, and $\{\Delta e_t\}$ should be stationary, and so the first difference of the regression equation

$$\Delta y_t = a_1 \Delta x_t + \Delta e_t$$

should yield meaningful results. This of course doesn't apply if one of the trends is deterministic, since in that case differencing won't make the time series stationary.

4. *Both $\{y_t\}$ and $\{x_t\}$ are non-stationary and integrated of the same order, but the sequence of residuals $\{e_t\}$ is stationary.* In this case, we say that $\{x_t\}$ and $\{y_t\}$ are "cointegrated." A trivial example would be that both time series are random walks, with the underlying white noise processes being perfectly correlated. Another example would be where $\{y_t\}$ and $\{x_t\}$ are defined by

$$y_t = \mu_t + \epsilon_{yt}$$

$$x_t = \mu_t + \epsilon_{xt}$$

where $\{\mu_t\}$ is itself a random walk process:

$$\mu_t = \mu_{t-1} + \epsilon_t$$

Both $\{y_t\}$ and $\{x_t\}$ are integrated of order 1, but $y_t - x_t = \epsilon_{yt} - \epsilon_{xt}$ is stationary.

14 Filters for trend and stochastic components

14.1 Hodrick-Prescott filter

Suppose we have some time series $\{y_t\}$, with $t = 1, 2, \ldots, T$. We can think of this series as being composed of a trend component $\{\tau_t\}$ and a cyclical component $\{c_t\}$, so that

$$y_t = \tau_t + c_t$$

The idea behind the Hodrick-Prescott filter is to estimate the trend component by finding $\{\tau_t\}$ such that

$$\sum_{t=1}^{T}(y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1}[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})]$$

is minimized. The first sum constrains the sum of squared differences $y_t - \tau_t$ (the cyclical component c_t). The second sum constrains the second differences of the trend component (essentially penalizing volatility in the trend); the constant λ can be adjusted to get a more or less smooth trend, with a higher value of λ corresponding to a smoother trend – Hodrick and Prescott recommended $\lambda = 1600$ for quarterly data. Once $\{\tau_t\}$ is estimated in this way, $\{c_t\}$ can then be calculated as $c_t = y_t - \tau_t$.

There are a number of problems with the Hodrick-Prescott filter. For example, if the process underlying $\{y_t\}$ is not integrated of order 2, then the estimate for $\{c_t\}$ can contain spurious cyclical fluctuations since the filter will force $\{c_t\}$ to be essentially a smoothed version of the second differences of $\{\tau_t\}$.

14.2 Hamilton filter

James Hamilton (UCSD) proposed an alternative to the Hodrick-Prescott filter that doesn't have those problems. Hamilton's idea is to simply estimate for each y_{t+h} the coefficients $\alpha, \beta_0, \beta_1, \ldots, \beta_p$ in the regression equation

$$y_{t+h} = \alpha + \beta_0 y_t + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + e_t$$

and then estimate the trend component as

$$\tau_{t+h} = \hat{y}_{t+h}$$

and the cyclical component as

$$c_{t+h} = y_{t+h} - \tau_{t+h} = \hat{e}_t.$$

The chosen value of h should correspond to the time horizon likely to be incorrectly predicted, and h and p should be multiples of the number of samples in a given year in the data being used. Hamilton recommended $h = 8$ and $p = 4$ for quarterly data.

14.3 Beveridge-Nelson decomposition

Beveridge and Nelson showed that it is possible to decompose any ARIMA$(p, 1, q)$ model into the sum of a trend component plus a stationary component.

The idea is based on "Wold's decomposition theorem," which states that any covariance-stationary time series $\{y_t\}$ can be written as the sum of a deterministic and a stochastic component. That is, if $\{y_t\}$ is covariance-stationary, then

$$y_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i} + \eta_t,$$

for some deterministic process $\{\eta_t\}$, white noise $\{\epsilon_t\}$, and [possibly] infinite vector of moving average coefficients $(b_0, b_1, b_2, \dots)$.

If $\{y_t\}$ is an ARIMA$(p, 1, q)$ process, then the first difference $\{\Delta y_t\}$ is stationary, so Wold's decomposition theorem implies that there is a C (of possibly infinite degree) and sequence $\{\eta_t\}$ such that

$$\Delta y_t = C(L)\epsilon_t + \eta_t.$$

Since 1 is a root of $C(L) - C(1)$, there is some polynomial C^* such that $C(L) - C(1) = C^*(L)(1 - L)$. So we can write

$$\begin{aligned} \Delta y_t &= [C^*(L)(1 - L) + C(1)]\epsilon_t + \eta_t \\ &= C^*(L)\Delta\epsilon_t + C(1)\epsilon_t + \eta_t. \end{aligned}$$

Thus

$$y_t = \underbrace{C^*(L)\epsilon_t}_{\text{stationary}} + \underbrace{C(1)\sum_{i=1}^{t}\epsilon_i + \sum_{i=1}^{t}\eta_i}_{\text{trend}}.$$

Practically speaking, to perform an approximate version of this decomposition on some time series data $\{y_t\}$ $(t = 1, \dots, T)$ without knowing the exact parameters of the time series, we would take the following steps:

1. Take the first difference to get $\{\Delta y_t\}$ and select the best-fitting ARMA(p, q) model for that sequence.

2. Using the best-fitting ARMA model, for each $t = 1, \dots, T$, find the h-step ahead forecasts for $h = 1, \dots, s$, and use those forecasts to construct the sum

$$\tau_t := \mathrm{E}_t[\Delta y_{t+s} + \Delta y_{t+s-1} + \cdots + \Delta y_{t+1}] + y_t.$$

(Note that s is a constant that depends on the particular ARIMA model we're dealing with. Beveridge and Nelson used $s = 100$ in their original work, but a much smaller value of s is sometimes appropriate depending on how fast the forecasts decay. What we're trying to do is find a reasonable approximation for $\lim_{s\to\infty} \mathrm{E}_t y_{t+s}$.)

3. We then treat the $\{\tau_t\}$ sequence as the trend component and construct the stationary component as

$$c_t := y_t - \tau_t = -\mathrm{E}_t[\Delta y_{t+s} + \Delta y_{t+s-1} + \cdots + \Delta y_{t+1}].$$

14.4 Theoretical Beveridge-Nelson decomposition example

Suppose that we want to find the Beveridge-Nelson decomposition of the ARIMA$(1, 1, 1)$ process

$$\Delta y_t = a_1 \Delta y_{t-1} + \epsilon_t + b_1 \epsilon_{t-1}.$$

We can write this as

$$\Delta y_t = a_1 L \Delta y_t + (1 + b_1 L)\epsilon_t$$

so

$$(1 - a_1 L)\Delta y_t = (1 + b_1 L)\epsilon_t$$

and

$$\Delta y_t = \frac{1 + b_1 L}{1 - a_1 L}\epsilon_t.$$

We can let $C(L) = (1 + b_1 L)/(1 - a_1 L)$ so that $\Delta y_t = C(L)\epsilon_t$. (This is a rational equation, but we can take its Taylor expansion to represent it as a infinite-degree polynomial equation, as in the explanation above.) We want to set

$$\begin{aligned} C^*(L) &= \frac{C(L) - C(1)}{1 - L} \\ &= \frac{\frac{1+b_1 L}{1-a_1 L} - \frac{1+b_1}{1-a_1}}{1 - L} \\ &= \frac{-a_1 - b_1 + a_1 L + b_1 L}{(1 - a_1)(1 - a_1 L)(1 - L)} \\ &= -\frac{a_1 + b_1}{(1 - a_1)(1 - a_1 L)} \end{aligned}$$

So

$$\begin{aligned}\Delta y_t &= C^*(L)(1-L)\epsilon_t + C(1)\epsilon_t \\ &= -\frac{a_1+b_1}{(1-a_1)(1-a_1L)}\Delta\epsilon_t + \frac{1+b_1}{1-a_1}\epsilon_t,\end{aligned}$$

and

$$y_t = -\frac{a_1+b_1}{(1-a_1)(1-a_1L)}\epsilon_t + \frac{1+b_1}{1-a_1}\sum_{i=0}^{\infty}\epsilon_{t-i}.$$

The first part is the stationary component, and the second part is the trend component. We can get an alternate form by multiplying out the $(1-a_1L)$ term and simplifying:

$$y_t = a_1y_{t-1} + \epsilon_t + (1+b_1)\sum_{i=1}^{\infty}\epsilon_{t-i}.$$

We can get the general solution by taking the Taylor series expansion of $1/(1-a_1L)$ so that we end up with

$$C^*(L) = -\frac{a_1+b_1}{1-a_1}\sum_{i=0}^{\infty}a_1^iL^i.$$

So we can write the general solution as

$$y_t = \underbrace{-\frac{a_1+b_1}{1-a_1}\sum_{i=0}^{\infty}a_1^i\epsilon_{t-i}}_{\text{stationary component}} + \underbrace{\frac{1+b_1}{1-a_1}\sum_{i=0}^{\infty}\epsilon_{t-i}}_{\text{stochastic trend}}.$$

We can verify that the left part is in fact stationary, since $|a_1| < 1$ in order for $\{\Delta y_t\}$ to be stationary, and thus $\sum_{i=0}^{\infty}|a_1|^i$ converges.

Let's assign values to the constants a_1 and b_1 from part (d):

$$\Delta y_t = 0.5\Delta y_{t-1} + \epsilon_t + 0.7\epsilon_{t-1}$$

Based on part (d), we can write

$$\begin{aligned}y_t &= -\frac{0.5+0.7}{(1-0.5)(1-0.5L)}\epsilon_t + \frac{1+0.7}{1-0.5}\sum_{i=0}^{\infty}\epsilon_{t-i} \\ &= \frac{-2.4}{1-0.5L}\epsilon_t + 3.4\sum_{i=0}^{\infty}\epsilon_{t-i},\end{aligned}$$

and the general solution is

$$\begin{aligned} y_t &= -\frac{0.5+0.7}{1-0.5}\sum_{i=0}^{\infty} 0.5^i \epsilon_{t-i} + \frac{1+0.7}{1-0.5}\sum_{i=0}^{\infty}\epsilon_{t-i} \\ &= -2.4\sum_{i=0}^{\infty} 0.5^i \epsilon_{t-i} + 3.4\sum_{i=0}^{\infty}\epsilon_{t-i}. \end{aligned}$$

The stationary component is

$$c_t = -2.4\sum_{i=0}^{\infty} 0.5^i \epsilon_{t-i},$$

and the trend component is

$$\tau_t = 3.4\sum_{i=0}^{\infty}\epsilon_{t-i}.$$

Notice that

$$\begin{aligned} \mathrm{E}[c_t] &= -2.4\sum_{i=0}^{\infty} 0.5^i \mathrm{E}[\epsilon_{t-i}] \\ &= -2.4\sum_{i=0}^{\infty} 0.5^i \times 0 \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(c_t) &= (2.4)^2\sum_{i=0}^{\infty} 0.5^{2i}\mathrm{Var}[\epsilon_{t-i}] \\ &= 5.76\frac{\sigma^2}{1-0.25} \\ &= 7.68\sigma^2, \end{aligned}$$

where σ^2 is the variance of the $\{\epsilon_t\}$ sequence.

The trend component is a random walk.

15 VAR and SVAR models

A VAR will take the form

$$x_t = A_0 + A_1 x_{t-1} + A_2 x_{t-2} + \cdots + A_p x_{t-p} + e_t$$

while an SVAR (structural VAR) will take the form

$$Bx_t = \Gamma_0 + \Gamma_1 x_{t-1} + \Gamma_2 x_{t-2} + \cdots + \Gamma_p x_{t-p} + \epsilon_t.$$

Given a structural VAR model, we can derive a VAR model by multiplying by the inverse of the B matrix.

The structural VAR is a more direct representation of the underlying "structural" relationships of the variables involved. Furthermore, the elements of the error vector ϵ_t (also called the "structural shocks") are uncorrelated, in contrast with the error terms e_t in the VAR model. The SVAR model also allows for contemporaneous effects of the variables on each other. However, those same feedback properties make inconsistent OLS estimates for the parameters in an SVAR model, making the reduced (VAR) model better for identifying the joint dynamics of the variables involved.

Given an SVAR model, the corresponding VAR is not unique; working backward from a given VAR model to an SVAR is not possible unless some simplifying assumptions are made.

15.1 Finding VAR from an SVAR

We start with the bivariate first-order SVAR model

$$\begin{aligned} y_t &= b_{10} - b_{12} z_t + \gamma_{11} y_{t-1} + \gamma_{12} z_{t-1} + \epsilon_{yt} \\ z_t &= b_{20} - b_{21} y_t + \gamma_{21} y_{t-1} + \gamma_{22} z_{t-1} + \epsilon_{zt} \end{aligned}$$

To derive a reduced-form VAR model, we first rearrange terms:

$$\begin{aligned} y_t + b_{12} z_t &= b_{10} + \gamma_{11} y_{t-1} + \gamma_{12} z_{t-1} + \epsilon_{yt} \\ b_{21} y_t + z_t &= b_{20} + \gamma_{21} y_{t-1} + \gamma_{22} z_{t-1} + \epsilon_{zt} \end{aligned}$$

We can then express this as

$$\begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix} \begin{pmatrix} y_t \\ z_t \end{pmatrix} = \begin{pmatrix} b_{10} \\ b_{20} \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{yt} \\ \epsilon_{zt} \end{pmatrix}$$

If we define

$$B = \begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix}, \quad x_t = \begin{pmatrix} y_t \\ z_t \end{pmatrix}, \quad \Gamma_0 = \begin{pmatrix} b_{10} \\ b_{20} \end{pmatrix}, \quad \Gamma_1 = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}, \quad \epsilon_t = \begin{pmatrix} \epsilon_{yt} \\ \epsilon_{zt} \end{pmatrix}$$

we can express the SVAR model as

$$Bx_t = \Gamma_0 + \Gamma_1 x_{t-1} + \epsilon_t.$$

To get a VAR model, we can just multiply both sides of the equation by B^{-1}:

$$x_t = A_0 + A_1 x_{t-1} + e_t,$$

where $A_0 = B^{-1}\Gamma_0$, $A_1 = B^{-1}\Gamma_1$, and $e_t = B^{-1}\epsilon_t$.

Using the notation

$$A_0 = \begin{pmatrix} a_{10} \\ a_{20} \end{pmatrix}, \quad A_1 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad e_t = \begin{pmatrix} e_{yt} \\ e_{zt} \end{pmatrix},$$

we can write the VAR model as

$$\begin{aligned} y_t &= a_{10} + a_{11} y_{t-1} + a_{12} z_{t-1} + e_{yt} \\ z_t &= a_{20} + a_{21} y_{t-1} + a_{22} z_{t-1} + e_{zt}. \end{aligned}$$

Inverting the B matrix, we get

$$B^{-1} = \frac{1}{1 - b_{12} b_{21}} \begin{pmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{pmatrix}$$

so

$$e_t = B^{-1}\epsilon_t = \frac{1}{1 - b_{12} b_{21}} \begin{pmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{pmatrix} \begin{pmatrix} \epsilon_{yt} \\ \epsilon_{zt} \end{pmatrix}$$

and thus we have

$$e_{yt} = \frac{\epsilon_{yt} - b_{12}\epsilon_{zt}}{1 - b_{12} b_{21}}, \quad e_{zt} = \frac{\epsilon_{zt} - b_{21}\epsilon_{yt}}{1 - b_{12} b_{21}}.$$

Since they are linear combinations of the white noise series ϵ_{yt} and ϵ_{zt}, the error terms e_{yt} and e_{zt} have mean zero, constant variance, and are individually serially uncorrelated. However,

$$\begin{aligned} Cov(e_{yt}, e_{zt}) &= E[e_{yt} e_{zt}] \\ &= E\left[\left(\frac{\epsilon_{yt} - b_{12}\epsilon_{zt}}{1 - b_{12} b_{21}} \right) \left(\frac{\epsilon_{zt} - b_{21}\epsilon_{yt}}{1 - b_{12} b_{21}} \right) \right] \\ &= \left(\frac{1}{1 - b_{12} b_{21}} \right)^2 E[(1 + b_{12} b_{21})\epsilon_{yt}\epsilon_{zt} - b_{21}\epsilon_{yt}^2 - b_{12}\epsilon_{zt}^2] \\ &= -\frac{b_{21}\sigma_y^2 + b_{12}\sigma_z^2}{(1 - b_{12} b_{21})^2}, \end{aligned}$$

where σ_y^2 is the variance of ϵ_{yt}, and σ_z^2 is the variance of ϵ_{zt}.

Suppose we force $b_{12} = 0$. We can see in the original SVAR model that this would mean that y_t has a contemporaneous effect on z_t, but z_t has no contemporaneous effect on y_t (i.e., z_t affects y_t only in the first lag). Based on the above discussion about e_t, this would give us

$$e_{yt} = \epsilon_{yt}, \quad e_{zt} = \epsilon_{zt} - b_{21}\epsilon_{yt}$$

so

$$Cov(e_{yt}, e_{zt}) = Cov(e_{zt}, e_{yt}) = E[\epsilon_{yt}(\epsilon_{zt} - b_{21}\epsilon_{yt})] = -b_{21}\sigma_y^2$$

and

$$\mathrm{Var}(e_{yt}) = \mathrm{E}[\epsilon_{yt}^2] = \sigma_y^2,$$

$$\begin{aligned}
\mathrm{Var}(e_{zt}) &= \mathrm{E}[(\epsilon_{zt} - b_{21}\epsilon_{yt})^2] \\
&= \mathrm{E}[\epsilon_{zt}^2] - 2b_{21}\mathrm{E}[\epsilon_{yt}\epsilon zt] + b_{21}^2\mathrm{E}[\epsilon_{yt}^2] \\
&= \sigma_z^2 + b_{21}^2\sigma_y^2
\end{aligned}$$

meaning that the covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_y^2 & -b_{21}\sigma_y^2 \\ -b_{21}\sigma_y^2 & \sigma_z^2 + b_{21}^2\sigma_y^2 \end{pmatrix}.$$